Genomics and Proteomics in Epidemiology

Treasure Trove or "High-Tech Stamp Collecting"?

David J. Hunter

Even a brief perusal of current epidemiology journals and meeting programs shows the vibrant life and potential of epidemiology in the 21st century. Even if no new epidemiologic methods or technologies were developed, we could still make major contributions to understanding the etiology of the many diseases that are underresearched, or are still poorly understood despite substantial research, or are newly emerging. There is still plenty to be achieved with the well-designed questionnaire, the optimization of response and follow-up rates, and the 2 × 2 table.

However, this is also a time of rapid development of technologies that will facilitate epidemiologic research. Epidemiologists have always been opportunistic in the adoption of new technologies for exposure assessment or outcome definition, and many new molecular tools have already become available in the last 25 years. In the early 1980s, Perera and Weinstein[1] extended the term "molecular epidemiology" from the study of molecular characterization of infectious organisms to the study of exposure and susceptibility relevant to cancer and other chronic diseases. For some exposure–disease relations, molecular methods of exposure assessment have identified the etiologic culprits in a rogues' gallery of suspects—such as the association of human papillomaviruses with cervical cancer, after observational studies had strongly implicated aspects of sexual behavior as increasing risk, and seroepidemiologic studies had suggested associations with a number of other sexually transmitted infections. Similarly, the identification of somatic mutations characteristic of exposure to aflatoxin in the *TP53* gene in liver tumors provided a crucial piece of evidence to bolster a previously disputed set of ecologic and dietary studies associating aflatoxin exposure with liver cancer risk.[2] Application of novel biomarkers as measures of exposure, or to assess the misclassification associated with other methods, will offer new opportunities to define exposure–disease relations. Much of the rest of this article discusses the scale and accuracy of new methods to assess between-person inherited genetic variation. However, we (and the people who fund our work) should not forget the need to better assess environmental exposures, because these are often modifiable and remediable, unlike our inherited genotypes.

In the last decade, the development of individual biomarkers has been revolutionized by "-omics" approaches. The suffix "-omic" is from the Greek meaning "all" or "every"; thus, genomics is the study of all of the genes in an organism. The first genomic technology in widespread use was the expression array, which enabled the study of the degree of expression of all known genes in an organism (sometimes referred to as the "transcriptome"). The major application in large-scale human studies has been the analysis of tumors. The ability to measure the transcriptome of individual tumors permitted the description of different patterns of variation in gene expression within tumor types previously thought to be a single type based on conventional histology. These patterns have been shown to be associated with prognosis[3] and treatment response.[4] For some tumors, the different patterns may derive from different cells of origin.[5]

Submitted 5 April 2006; accepted 19 May 2006.

From the Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, MA; and Channing Laboratory, Brigham and Women's Hospital, Boston, MA.

Editors' note: *A related article appears on page 490.*

Address correspondence to: David J. Hunter, Department of Epidemiology, 677 Huntington Ave., Boston, MA 02115. E-mail: dhunter@hsph.harvard.edu.

ISSN: 1044-3983/06/1705-0487
DOI: 10.1097/01.ede.0000229955.07579.f0

What has not been explored in any depth is whether these patterns may have different etiologies. Our relative failure to understand the causes of lymphomas, for instance, may be due in part to having collapsed a number of different etiologic entities as a single disease with resulting misclassification and loss of power. The application of the new technologies was initially limited by the need for fresh-frozen tumors as sources of RNA—tissues that only relatively few clinical investigators had available, almost always from samples stripped of identifiers or with only very limited risk-factor information. Newer techniques condense the genomic analyses to a smaller number of predictors of these patterns that can be measured in paraffin-embedded tissues. This technologic change should enable these analyses to be applied to large-scale epidemiologic studies.

As a community, however, we are still in the earliest stages of being able to apply these technologies, and the consequences for the sample sizes needed to adequately explore the etiology of tumor subtypes are daunting. Studies powered for a cancer type as if it is a single entity will obviously need to be much larger if the cancer is, in fact, composed of multiple etiologic entities. Replication of results will also demand larger studies, but also agreement on the definition of the subtypes. It may be possible to extend this paradigm that has emerged for cancers to other diseases for which affected tissues can be sampled.

Technologies have developed to the point at which the genome-wide assessment of inherited DNA variation is now possible. Confident predictions have been made that, in 5 to 10 years, it will be possible to determine the sequence of an individual's genome in a matter of days for several thousands of dollars (compared with many years and many hundreds of millions of dollars for the first consensus human genome sequence). It has turned out, however, that most humans differ from another randomly chosen human at only approximately one in every 1,250 nucleotides[6] with only approximately 10 million such variants existing in the genome in the populations with a frequency above 1%.[7] Correlation, or linkage disequilibrium between these single nucleotide polymorphic (SNP) variants, means that a set of approximately 500,000 SNPs may be adequate to describe much of the between-person differences in the human genome sequence.

Whole-genome SNP scans are now feasible at a cost of $1000 per sample, and whole-genome scans are underway or being planned for many diseases. The scans will identify SNPs that are more (or less) common in cases than in controls; most of these SNPs will not be the causal alleles, but may be markers of them. Subsequent fine mapping will be required to identify which gene variants are likely to be causal by virtue of their location and predicted function. Further laboratory analyses will often be necessary to obtain proof of function, and genetically engineered animals may be required to convert this proof of function into proof of physiological relevance.

Using these methods, it is almost certain that many of the causal gene variants that underlie susceptibility or resistance to common diseases will be found in the next few years. Classic methods of genetic epidemiology such as twin studies have defined a spectrum of inherited contribution to diseases ranging from minimal to very high.[8] Thus, over the next few years, it is likely that a substantial proportion of the unexplained causation of many of these diseases will be determined, leading to a burst of etiologic understanding to rival the golden age of infectious disease epidemiology in the latter stages of the 19th century. Whether the information leads to a similar burst of use in public health or clinical interventions remains to be seen.

The basic dogma of molecular biology is "DNA makes RNA makes protein," and proteomics is the next "-omic" technology that will impact the practice of epidemiology. Proteomics describes the simultaneous assessment of large numbers of proteins and protein fragments in substrates that may include blood, urine, and tissue samples. The first major application is likely to be in the detection and early diagnosis of disease. Again, most applications have been to cancer with a prominent first report that a distinctive pattern of low-molecular-weight proteins characterizes the sera of women with early-stage ovarian cancer and is very rarely present in control sera.[9] This study has proved controversial, and commentators[10] have offered alternate interpretations of the initial observations. There is the possibility that the basic epidemiologic principle of treating samples from cases identically to those from controls was not followed, with the apparent differences between cases and controls reflecting artifacts of sample handling. Nonetheless, it is likely that the science of screening for disease will be informed by better proteomic assays assessed in more rigorous studies. In addition to the public health impact of the potential for earlier disease diagnosis, these technologies may permit the epidemiologic study of early-stage disease in a manner not possible when we relied on clinically apparent, often advanced disease.

Another application of proteomics is the assessment of normal and abnormal protein patterns not directly connected to a disease diagnosis, sometimes referred to as "metabolomics." In recent years, the emergence of the "metabolic syndrome" and "syndrome X," and the interest in inflammation as a common pathway to many diseases, have led to the epidemiologic study of these intermediate metabolic conditions as a legitimate option. Metabolomics may assist in defining patterns of protein abundance that have correlations with risk of future diseases, and thus may provide options for outcome assessment that are not centered on the classic pathologic classifications of disease. Again, such a development would help move the study of disease development to a preclinical stage, an idea that has parallels in the enthusiasm for the study of "early life determinants" of diseases. The chief problem for this area has been the long latency between exposures in early life and disease development in middle or late life and the consequent difficulty in assembling data sets with adequately classified exposure information. Metabolomics may help provide a bridge between the early life exposures and later diseases.

Having briefly surveyed some of the exciting potential of "-omic" science applied to epidemiology, some caveats are in order. The scale of data quantity in these measurements dwarfs the information per subject we typically acquire. Implicit in the "-omic" concept, with its simultaneous measurement of tens or hundreds or thousands of parameters, is the idea that only a tiny fraction of the information is likely to have biologic relevance to the end point under study. Sorting out this signal from noise has required the rapid development of "data-mining" algorithms that must then be subjected to "test–retest" procedures to determine their reproducibility.

All of this is a long way from the "hypothesis-driven" testing that motivates classic frequentist approaches to determining levels of statistical significance. The tension is alive and well between the descriptive approaches inherent in the analysis of these data sets and the drive for "prespecifying" hypotheses and limiting multiple comparisons in both observational and interventional human studies. In many ways, the philosophy of "test–retest," so central to the "-omic" analyses, is akin to the concept in observational epidemiology that reproducibility of findings is the key to interpretation of data. This philosophy moves us away from the "one definitive study" model of epidemiologic research that has led to a barrage of false-positives entering both the literature and the popular press. Indeed, the massive amounts of data involved, and the potentially incompatible measurements across platforms, raise new problems for reproducibility. Consortia are being formed in an attempt to standardize methods before studies are performed so they can be compared or pooled.[11] This development should accelerate the assessment of reproducibility and diminish the specter of publication bias.

Epidemiologists should not be seduced by technology into measuring observations for the sake of doing so. Tony McMichael[12] warned us against this when he wrote about the dangers of "high-tech stamp collecting." Epidemiology has been described as "the art of the possible," but there is a difference between the study of the important diseases that are possible to study and the study of what is possible just for the sake of doing so. "Because it is there" (Mallory's answer to the question of why he wanted to climb Everest) is not a good mantra for the epidemiologist. We have to learn what technologies can be usefully applied to which questions rather than applying everything to anything. Inevitably, a certain amount of "stamp collecting" may be required to learn the use (or nonuse) of new technologies. Still, the principles of good study design and adequate sample size must not be forgotten in any drive to generate novel data.

In addition, the lure of new technologies should not distract us from following social, economic, demographic, or ecologic explanations for disease etiology. Even here, "-omic" technologies may have much to offer; for example, the relation of serotonin transporter polymorphisms with risk of depression associated with stressful life events may help clarify the complex etiology of this disease.[13]

Finally, the application of these technologies to epidemiology poses challenges to the organization of our science. We all face the challenge of training (and retraining) in these new areas, and we try to keep abreast of an apparently exponentially increasing set of opportunities while still paying close attention to the rigor of epidemiologic study design and analyses. These latter concerns can appear pedantic and arcane to the laboratory investigator accustomed to using "freezer controls" or to those with a mindset that only very large differences are interesting. Again, tension exists between the possibility that a new technology may permit the discovery of a strong risk factor that can be detected in "quick and dirty" studies and the necessity to minimize bias through painstaking attention to detail to be as certain as possible about the weaker factors. Epidemiology will be well served by staying the course on optimization of study design so that, at a minimum, these studies can sort the wheat from the chaff of "quick and dirty" findings.

All of this has to be done in the context of social beliefs that often trend toward genetic determinism augmented in recent years by the publicity and hoopla surrounding the "race" to decipher the human genome. As epidemiologists, we are aware that most diseases are due to the complex interplay of genes and environment. We know that the public health impact of avoidable lifestyle and environmental risk factors is far from fully achieved. A public focus on the concept of inherited disease susceptibility or resistance could lead to a diminution of our ability to provide "broad brush" or "one size fits all" lifestyle advice, currently the mainstay of public health campaigns. However, this and other dangers (eg, the potential misuse of genetic information for inappropriate screening) are all good reasons for us to get involved. Pandora's box is being opened, the genie is escaping the bottle, and we must seize the opportunity to harness the magic for public health good.

ABOUT THE AUTHOR

DAVID J. HUNTER is the Vincent L. Gregory Professor in Cancer Prevention at the Harvard School of Public Health. He heads the Program in Molecular and Genetic Epidemiology and is Project Director for the Nurses Health Study II. He is also director of the Polymorphism Detection Core, which provides genotyping facilities for investigators at HSPH and elsewhere. His research includes studies of polymorphisms in carcinogen-metabolizing genes and breast and colon cancer.

REFERENCES

1. Perera FP, Weinstein IB. Molecular epidemiology and carcinogen-DNA adduct detection: new approaches to studies of human cancer causation. *J Chronic Dis*. 1982;35:581–600.
2. Hsu IC, et al. Mutational hotspot in the p53 gene in human hepatocellular carcinomas. *Nature*. 1991;350:427–428.
3. Pusztai L, et al. Clinical application of cDNA microarrays in oncology. *Oncologist*. 2003;8:252–258.
4. Ramaswamy S, Golub TR. DNA microarrays in clinical oncology. *J Clin Oncol*. 2002;20:1932–1941.
5. Sorlie T, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*. 2003;100:8418–8423.
6. Reich DE, et al. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet*. 2002;32:135–142.
7. Altshuler D, Clark AG. Genetics. Harvesting medical information from the human family tree. *Science*. 2005;307:1052–1053.
8. Chakravarti A, Little P. Nature, nurture and human disease. *Nature*. 2003;421:412–414.
9. Petricoin EF, et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*. 2002;359:572–577.
10. Diamandis EP. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *J Natl Cancer Inst*. 2004;96:353–356.
11. Hunter DJ, et al. A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. *Nat Rev Cancer*. 2005;5:977–985.
12. McMichael AJ. Invited commentary—'molecular epidemiology': new pathway or new travelling companion? *Am J Epidemiol*. 1994;140:1–11.
13. Caspi A, et al. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science*. 2003;301:386–389.